A statistical approach for inferring the three-dimensional structure of the genome

Nelle Varoquaux, Ferhat Ay, William Stafford Noble, Jean-Philippe Vert

A statistical approach for inferring the three-dimensional structure of the genome

Nelle Varoquaux [1,2,3], Ferhat Ay [4], William Stafford Noble [4,5*], and Jean-Philippe Vert [1,2,3*]

[1] Centre for Computational Biology, Mines ParisTech, Fontainebleau, F-77300 France
[2] Institut Curie, Paris, F-75248, France
[3] U900, INSERM, Paris, F-75248, France
[4] Department of Genome Sciences, University of Washington, Seattle, WA, USA, 98195.
[5] Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA, 98195

Abstract

Motivation: Recent technological advances allow the measurement, in a single Hi-C experiment, of the frequencies of physical contacts among pairs of genomic loci at a genome-wide scale. The next challenge is to infer, from the resulting DNA-DNA contact maps, accurate three dimensional models of how chromosomes fold and fit into the nucleus. Many existing inference methods rely upon *multidimensional scaling* (MDS), in which the pairwise distances of the inferred model are optimized to resemble pairwise distances derived directly from the contact counts. These approaches, however, often optimize a heuristic objective function and require strong assumptions about the biophysics of DNA to transform interaction frequencies to spatial distance, thereby leading to incorrect structure reconstruction.

Methods: We propose a novel approach to infer a consensus three-dimensional structure of a genome from Hi-C data. The method incorporates a statistical model of the contact counts, assuming that the counts between two loci follow a Poisson distribution whose intensity decreases with the physical distances between the loci. The method can automatically adjust the transfer function relating the spatial distance to the Poisson intensity and infer a genome structure that best explains the observed data.

Results: We compare two variants of our Poisson method, with or without optimization of the transfer function, to four different MDS-based algorithms—two metric MDS methods using different stress functions, a nonmetric version of MDS, and ChromSDE, a recently described, advanced MDS method—on a wide range of simulated datasets. We demonstrate that the Poisson models reconstruct better structures than all MDS-based methods, particularly at low coverage and high resolution, and we highlight the importance of optimizing the transfer function. On publicly available Hi-C data from mouse embryonic stem cells, we show that the Poisson methods lead to more reproducible structures than MDS-based methods when we use data generated using different restriction enzymes, and when we reconstruct structures at different resolutions.

1 Introduction

Spatial and temporal three-dimensional (3D) genome architecture is thought to play an important role in many genomic functions, but is still poorly understood (van Steensel and Dekker, 2010). In recent years, the technique of chromosome conformation capture (3C) (Dekker et al., 2002), which identifies physical contacts between different genomic loci and yields information about their relative spatial distance in the nucleus, has paved the way for the systematic analysis

*to whom correspondence should be addressed: william-noble@uw.edu, jean-philippe.vert@mines.org

of the 3D structure of DNA. Coupled with high-throughput sequencing, genome-wide conformation capture assays, broadly referred to as *Hi-C* (Lieberman-Aiden et al., 2009), have emerged as promising techniques to investigate the global structure of DNA at various resolutions. Hi-C has opened new avenues to understanding many biological processes including gene regulation, DNA replication, somatic copy number alterations and epigenetic changes (Shen et al., 2012; Ryba et al., 2010; De and Michor, 2011; Dixon et al., 2012).

A typical Hi-C experiment yields a DNA *contact map*, that is, a matrix indicating the frequency of interactions between all pairs of loci at a given resolution. A fundamental question is then to reconstruct the 3D structure of the genome from this contact map. Two general approaches have been proposed for that purpose: (i) *consensus methods* that aim at inferring a unique mean structure representative of the data and (ii) *ensemble methods* that yield a population of structures.

Consensus approaches (Duan et al., 2010; Tanizawa et al., 2010; Bau et al., 2011) model each chromosome by a chain of beads, convert the contact map frequencies into pairwise distances (which we refer as *wish distances*) using various biophysical models of DNA, and infer a 3D conformation that best matches the pairwise distances by solving a multidimensional scaling (MDS) problem (Kruskal and M., 1977). Converting interaction counts to physical wish distances requires, however, strong assumptions which are not always met in practice. For example, this mapping may change from one organism to another (Fudenberg and Mirny, 2012), from one resolution to another (Zhang et al., 2013), or from one time point to another during cell cycle (Le et al., 2013). To alleviate this problem, Zhang et al. (2013) proposed ChromSDE, a method that jointly optimizes the 3D structure and a parameter of the function that maps contact frequencies to spatial distances, in addition to modifying the objective function of MDS. Ben-Elazar et al. (2013) proposed an approach akin to *nonmetric MDS* (Kruskal, 1964), where the 3D structure and the wish distances are alternatingly optimized in an attempt to preserve coherence between the ranking of pairwise distances and the ranking of pairwise contact frequencies.

As for the ensemble methods, Rousseau et al. (2011) and Hu et al. (2013) describe two formal probabilistic models of contact frequencies and their relationship with physical distances. They then use a Markov chain Monte Carlo (MCMC) sampling procedure to produce an ensemble of 3D structures consistent with the observed contact counts. Kalhor et al. (2011) propose an optimization framework that generates a population of structures by enforcing each contact to define an active constraint in only a fraction of the inferred structures, thereby mimicking the heterogeneity of contacts coming from each cell in the Hi-C sample. Applying a similar method to budding yeast, Tjong et al. (2012) demonstrate that a large population of structures inferred using known physical constraints of yeast genome architecture can recapitulate, to a large extent, the consensus contact map observed from Hi-C experiments.

Both consensus and ensemble models have benefits and limitations. Ensemble approaches are biologically more accurate, because Hi-C data is derived from a population of cells, each with potentially a unique 3D architecture. An inferred population of 3D structures may therefore better reflect the diversity of structures than a single consensus structure. In concordance with such ensemble methods, a recent development in Hi-C technology, assaying chromatin conformation at a single cell level, demonstrates that chromatin structure varies highly from cell to cell by modeling the single-copy X chromosomes of a male mouse cell line (Nagano et al., 2013).

However, an ensemble approach raises the question of interpretability: one often has to fall back to interpreting a mean signal from the population structure (Kalhor et al., 2011) or to selecting a few structures, representative in some way of the diversity of the population (Rousseau et al., 2011). Consensus methods, in contrast, provide a single structure more amenable to visual inspection and analysis. This structure can be seen as a useful *model* to recapitulate the rich information captured in Hi-C data and to allow easy integration with other sources of data,

such as RNA-seq, which are usually also population based. In addition, despite the stochasticity of cell-to-cell variations, certain hallmarks of genome organization observed by consensus methods, such as chromosome territories or topological domain organization, are conserved across different cells (Nagano et al., 2013; Hu et al., 2013). Computationally, ensemble methods are more demanding than consensus methods since they need to sample from a very large dimensional space of possible structures with complicated likelihood landscapes. Optimization-based consensus methods are usually faster to converge to a local optimum, but may miss the global optimum corresponding to the best structure when the objective function is non-convex.

In this work, we focus on the consensus approach, and we propose a new method to infer a 3D structure from Hi-C data. We propose to replace the arbitrary loss function minimized by existing MDS-based approaches by a better-motivated likelihood function derived from a statistical model, similar to the one use by a previous ensemble method (Hu et al., 2013). Specifically, our proposed method models the interaction frequency between two loci by a Poisson model (PM), the intensity of which decreases with the increasing spatial distance between the pair of loci. Similar to the problem of inferring the wish distances from interaction frequencies faced by MDS-based approaches, our model faces the difficulty of transforming spatial distances into intensities of the Poisson distribution. To solve this problem, we propose two variant methods. The first method (PM1) uses a default transfer function motivated by a biophysical model, whereas the second method (PM2) uses a parametric family of transfer functions, the parameters of which are automatically optimized together with the 3D structure to best explain the observed data.

We compare both PM variants to four MDS-based methods, including metric MDS with two stress functions, nonmetric MDS and ChromSDE. We demonstrate on simulated data that the new models reconstruct more accurate 3D structures than all MDS-based methods, especially at low coverage and high resolution. We also assess the negative effect of using an incorrect transfer function, and we show that PM2 is able to overcome this difficulty. On real data, we show that, compared to MDS-based methods, PM1 and PM2 generate more similar models when applied to replicate experiments performed with different restriction enzymes or when applied to the same data at varying resolutions. The results suggest that the Poisson model methods we describe here provide promising alternatives to current methods for consensus DNA structure inference.

2 Approach

We model chromosomes as series of beads in 3D, each bead representing a genomic window of a given length, and we denote by $\mathbf{X} = (x_1, \ldots, x_n) \in \mathbb{R}^{3\times n}$ the coordinate matrix of the structure, where n denotes the total number of beads in the genome (for example, $n = 4457$ at 10kb resolution for the yeast genome) and $x_i \in \mathbb{R}^3$ represents the 3D coordinate of the i-th bead. The Hi-C data can be summarized as an n-by-n matrix $\mathbf{c}$ in which each row and column corresponds to a genomic locus, and each matrix entry c_{ij} is a number, called the *contact frequency* or *contact count*, indicating the number of times locus i and j were observed to contact one another. The matrix is by construction square and symmetric.

2.1 Data normalization

The raw contact count matrix suffers from many biases, some technical (from the sequencing and mapping) and others biological (inherent to the physical properties of chromatin) (Yaffe and Tanay, 2011; Imakaev et al., 2012). Therefore, before inferring the 3D structure of the genome, we normalize each raw contact matrix using iterative correction and eigenvalue decomposition (ICE) (Imakaev et al., 2012), a method based on the assumption that all loci should interact equally. Due to mappability issues, some beads have zero contact counts. We remove these

beads from the optimization and only try to infer the positions of beads with nonzero contact counts.

2.2 MDS-based methods

2.2.1 Metric MDS

Metric MDS is a classical method to infer coordinates of points given their approximate pairwise Euclidean distances (Kruskal and M., 1977). To use MDS in the context of DNA structure inference from Hi-C data, we need to assign each pair of beads (i, j) a physical wish distance δ_{ij}—i.e., the distance that we aim to capture with our 3D model—derived from the bead pair's contact count c_{ij}. Performing this assignment requires us to decide how contact counts are transformed into physical distances. In Section 2.4 we discuss a commonly used transformation of the form $\delta_{ij} = \gamma c_{ij}^{-3}$ if $c_{ij} > 0$ motivated by polymer physics. Metric MDS then places all the beads in 3D space such that the Euclidean distance $d_{ij}(\mathbf{X}) = \|x_i - x_j\|$ between the beads i and j is as close as possible to the wish distance δ_{ij}. Denoting by $\mathcal{D}$ the subset of indices whose distances we wish to constrain (typically, the set of pairs (i, j) with non-zero contact counts $c_{ij} > 0$), metric MDS attempts to minimize the following objective function, usually called the *raw stress*:

$$\underset{\mathbf{X}}{\text{minimize}} \qquad \sum_{(i,j)\in\mathcal{D}} \big(d_{ij}(\mathbf{X}) - \delta_{ij}\big)^2 . \tag{1}$$

In two previous studies that use metric MDS, Duan et al. (2010) and Tanizawa et al. (2010) infer the 3D structure of DNA from Hi-C data by solving Equation 1, limiting $\mathcal{D}$ to pairs of indices with statistically significant contact counts (FDR 0.01%). Both methods use additional constraints such as minimum and maximum distances between adjacent beads, minimum pairwise distances between arbitrary beads to avoid clashes, and organism-specific constraints that concern the positioning of centromeres, telomeres and ribosomal RNA coding regions. In the experiments we present here, we simply solve Equation 1 without any constraints but including all pairs of beads with positive counts in $\mathcal{D}$, and we call the resulting method MDS1. In general, we have observed that adding constraints related to minimal and maximal distances between beads is unnecessary, because the structures found by MDS1 typically fulfill all of these constraints (data not shown).

A drawback of the raw stress of Equation 1 in our context is that the quadratic form is dominated by large values, corresponding to pairs of loci with large wish distances (i.e., small contact counts). Because these counts are less reliable than large contact counts, we propose a variant of MDS1, which we call MDS2, where we weight the contribution of a pair (i, j) in the stress by a factor inversely proportional to the square wish distance between the corresponding beads:

$$\underset{\mathbf{X}}{\text{minimize}} \qquad \sum_{(i,j)\in\mathcal{D}} \delta_{ij}^{-2}\big(d_{ij}(\mathbf{X}) - \delta_{ij}\big)^2 . \tag{2}$$

While other weighting schemes could be proposed to decrease the influence of pairs with large wish distances, we found this formulation to be quite robust in practice. Notice that MDS2 can be thought of as a quadratic approximation of the raw stress (minimized by MDS1) applied to log-transformed distances, because in the setting $d_{ij}(\mathbf{X}) \approx \delta_{ij}$ it holds that:

$$\begin{aligned}\sum_{(i,j)\in\mathcal{D}} \left(\log d_{ij}(\mathbf{X}) - \log \delta_{ij}\right)^2 &= \sum_{(i,j)\in\mathcal{D}} \log\left(\frac{d_{ij}(\mathbf{X})}{\delta_{ij}}\right)^2 \\ &\approx \sum_{(i,j)\in\mathcal{D}} \left(\frac{d_{ij}(\mathbf{X})}{\delta_{ij}} - 1\right)^2 .\end{aligned}$$

Both MDS1 and MDS2 implicitly ignore non-interacting pairs of beads (i.e., pairs with zero contact counts).

In addition to MDS1 and MDS2, we include in our benchmark ChromSDE (Zhang et al., 2013), a recently proposed method which also attempts to minimize a weighted stress function penalized by an additional term to push non-interacting pairs far from each other. In addition, ChromSDE optimizes the exponent of the transfer function from contact counts to wish distances.

2.2.2 Nonmetric MDS (NMDS)

The derivation of the transfer function from contact counts to 3D wish distances, needed by metric MDS-based methods, relies on strong assumptions about the physics of DNA (Section 2.4). NMDS (Shepard, 1962; Kruskal, 1964) offers an alternative way to proceed, which was proposed in the context of DNA structure inference from Hi-C data by Ben-Elazar et al. (2013). Instead of inferring physical distances from the contact matrices, NMDS relies on the sole hypothesis that if two loci i and j are observed to be in contact more often than loci k and ℓ, then i and j should be closer in 3D space than k and ℓ. Using this hypothesis, NMDS attempts to solve the following problem:

Problem 2.1 *Given a set of similarities or dissimilarities c_{ij}, find $\mathbf{X} \in R^{3\times n}$ such that:*

$$c_{ij} \geq c_{k\ell} \Leftrightarrow \|x_i - x_j\|_2 \leq \|x_k - x_\ell\|_2 \tag{3}$$

Equation 3 is known as the nonmetric constraint, or the ordinal constraint. This problem was first introduced by Shepard (1962) and formalized as an optimization problem by Kruskal (1964). It can be solved by minimizing the cost function:

$$\underset{\mathbf{X},\boldsymbol{\Theta}}{\text{minimize}} \sum_{i,j} \frac{\left(\|x_i - x_j\|_2 - \Theta(c_{ij})\right)^2}{\Theta(c_{ij})^2}, \tag{4}$$

with respect to the embedding $\mathbf{X}$ and the function Θ, where Θ is a decreasing function. Algorithms to solve this optimization problem involve iterating over two steps: (1) fixing Θ and minimizing the objective function with respect to $\mathbf{X}$ (hence falling back to solve MDS2), and (2) fitting Θ to the new configuration $\mathbf{X}$ subject to the ordinal constraints. This second step of the algorithm can be performed using an isotonic regression method, such as the pool adjacent violator algorithm (Best et al., 1999).

A trivial solution of this problem is to set Θ equal to 0. In this case all points will collapse on the origin. To avoid this collapse, we add additional constraints on $\mathbf{X}$ or on Θ, such as $\sum_{i,j} \|x_i - x_j\|_2 = K$ for some constant value of K.

2.3 Poisson model

Instead of metric or non metric MDS-based methods, which attempt to minimize a stress function that measures a discrepancy between the wish distances and the 3D distances of the structure, we propose to cast the problem of structure inference as a maximum likelihood problem. For that purpose, we need to define a probabilistic model of contact counts parametrized by the 3D structure that we want to infer from contact count observations.

For that purpose, we take a model similar to the one used in the BACH algorithm (Hu et al., 2013) and model the contact frequencies $(c_{ij})_{(i,j)\in\mathcal{D}}$ as independent Poisson random variables, where the Poisson parameter of c_{ij} is a decreasing function of $d_{ij}(\mathbf{X})$ of the form $\beta d_{ij}(\mathbf{X})^\alpha$, for some parameters $\beta > 0$ and $\alpha < 0$. We can then express the likelihood as

$$\ell(\mathbf{X}, \alpha, \beta) = \prod_{i,j} \frac{(\beta d_{ij}^\alpha)^{c_{ij}}}{c_{ij}!} \exp(-\beta d_{ij}^\alpha)\,.$$

By maximizing the log likelihood, a new optimization problem naturally emerges from this formulation:

$$\max_{\alpha,\beta,\mathbf{X}} \quad \mathcal{L}(\mathbf{X},\alpha,\beta) = \sum_{i<j\leq n} c_{ij}\alpha \log d_{ij} + c_{ij}\log\beta - \beta d_{ij}^{\alpha} \tag{5}$$

With this new formulation, we can either provide the parameter α, using prior knowledge, and only optimize the structure and β (which depends on the dataset), or we can use a nonmetric approach, by inferring α. We refer to the former as PM1 and to the latter as PM2.

PM2 is solved using a coordinate-descent algorithm: first choose randomly an $\mathbf{X}$ configuration, then iterate between maximizing $\mathcal{L}$ with respect to α and β and, fixing α and β and maximizing $\mathcal{L}$ with respect to $\mathbf{X}$. In this work, we try to initialize $\mathbf{X}$ with a good approximation of the solution by first evaluating the parameters α and β using some prior knowledge and initialize $\mathbf{X}$ with the inferred structure from the MDS.

All optimization problems (MDS1, MDS2, NMDS, PM1 and PM2) were solved using IPOPT, an interior point filter algorithm (Wächter and Biegler, 2006) and the isotonic regression implementation from the Python toolbox Scikit-Learn for NMDS (Pedregosa et al., 2011).

2.4 Default contact-to-distance transfer function

A prerequisite for both the MDS and the PM1 model (and for good initialization of the NMDS and PM2 methods) is a function that converts from contact counts to wish distances. Extensive previous studies of the behaviour of polymers in general and DNA in particular have yielded proposed relationships between, on the one hand, the genomic distance s and contact counts c and, on the other hand, genomic distance s and physical distances d for several classes of polymers (Grosberg et al., 1988; Lieberman-Aiden et al., 2009; Fudenberg and Mirny, 2012). For a fractal globule polymer, representative of mammalian DNA, the contact counts is inversely proportional to the genomic distance ($c \sim s^{-1}$), whereas the volume scales linearly with the subchain length ($d^3 \sim s$), from which we deduce a relationship between d and c of the form $d \sim c^{-1/3}$. For an equilibrium globule, representative of a smaller genome such as *S. cerevisae*, the relationships differ: $c \sim s^{-3/2}$ and $d \sim s^{1/2}$ up to a maximum distance, corresponding to the size of the nucleus in which the DNA is confined. Conveniently, coupling those two relationships for either type of polymer yields the same mapping between contact counts and physical distances:

$$d \sim c^{-1/3}. \tag{6}$$

Thus, by default we convert contact counts c_{ij} into 3D wish distances δ_{ij} using the following relationship

$$\delta_{ij} = \gamma c_{ij}^{-1/3}, \tag{7}$$

where γ defines the scale of the structure. In practice, we will not infer γ for the MDS and NMDS problem: the structures can easily be rescaled after convergence to match biological knowledge of the organism studied.

2.5 Data

In order to test various 3D architecture inference methods, we conducted experiments on both simulated datasets and publicly available genome-wide Hi-C datasets.

For the simulation, we generated 170 data sets using the yeast genome architecture proposed by Duan et al. (2010). Because the repetitive rDNA on yeast chromosome XII cannot be observed in practice, we discard all contacts involving these loci, and we do not infer the position of the corresponding rDNA. We generate these 170 datasets using the following model:

$$c_{ij} = P(\beta d_{ij}^{\alpha}), \tag{8}$$

where $\alpha = -3$ (corresponding to the theoretical exponent discussed in Section 2.4) and β varies between 0.01 and 0.7 (0.01, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7) with 10 different random generator seeds, thus obtaining 10 different datasets per parameter. The β parameter controls the number of contact counts in the datasets. A low β will yield a dataset with few counts; hence, the corresponding wish distance matrix will be less likely to be close to the true distance matrix. To estimate how noisy the generated data is, we compute the following measure of signal-to-noise ratio:

$$SNR = \frac{\sum c_{ij}}{\sqrt{\sum (\beta d_{ij}^{\alpha} - c_{ij})^2}} \,. \tag{9}$$

The numerator (the signal) corresponds to the number of counts, and the denominator (the noise) corresponds to the sum of deviation between each count and its expected value. We use this first ensemble of simulated datasets to assess the robustness to noise of the different methods. Note that in actual data, the SNR gets smaller when we sequence fewer reads or when we infer a structure at a higher resolution.

We simulated another ensemble of datasets to compare nonmetric and metric methods when the parameters provided to the different algorithms are not the correct ones. We generate 20 datasets according to Equation 8, with α between -4 and -2 (-4, -3.5, -3, -2.5, -2) and β between 0.4 and 0.7 (0.4, 0.5, 0.6, 0.7).

We also applied our methods to publicly available Hi-C data from mouse embryonic stem cells (mESC) (Dixon et al., 2012). We started with the data at 20 kb resolution and considered only chromosomes 1 to 19, with both available restriction enzymes (HindIII and NcoI). We then subsampled the data at resolutions of 100 kb, 200 kb, 500 kb and 1 Mb.

2.6 Structure similarity measures

In order to assess the ability of a method to reconstruct a known structure from simulated data, or the stability of the reconstructed structure with respect to change in resolution or library preparation, we need quantitative measures of similarity between 3D structure. We use two such measures: the root mean square deviation (RMSD) and the distance error, which we now explain.

The RMSD is a standard way to compare two sets of structures described by their coordinates $\mathbf{X}, \mathbf{X}' \in R^{3\times n}$, widely used for example to compare protein 3D structures. It is given by:

$$RMSD = \min_{\mathbf{X}^*} \sqrt{\sum_{i=1}^{n} (\mathbf{X}_i - \mathbf{X}_i^*)^2}$$

where the structure $\mathbf{X}^*$ is obtained by translating, rotating and rescaling $\mathbf{X}'$ ($\mathbf{X}^* = s\mathbf{R}\mathbf{X}' - \mathbf{t}$, where $\mathbf{R} \in R^{3\times 3}$ is a rotation matrix, $\mathbf{t} \in R^3$ is a translation vector, and s is a scaling factor). Because ChromSDE does not infer the relative position of chromosomes, the RMSD we report below are sums of RMSD computed independently on each chromosome.

We also directly compare the 3D distance matrices corresponding to the two structures with the distance error:

$$\text{distanceError} = \sqrt{\sum_{i,j=0}^{n} (d_{ij}(\mathbf{X}) - d_{i,j}(\mathbf{X}'))^2}$$

The main difference between the optimization formulated by ChromSDE and those of the other methods is the penalty assigned to non-interacting beads. Due to this penalty, ChromSDE should recover better long distances than other MDS-based methods. This property is not well captured by the RMSD measure, therefore, we also compute how well the distance matrix is

recovered with the distance error, which assigns most of the weight to long distances. We expect that methods based on MDS, which optimize an objective function based on the distance matrix, should perform better on this measure than others.

3 Results

To assess the relative strength of our new Poisson model-based methods, PM1 and PM2, we compare them to a panel of four MDS-based methods: MDS1, MDS2, NMDS and ChromSDE on simulated and real data.

3.1 Simulated Hi-C data

We first tested the six methods on data simulated as explained in Section 2.5.

3.1.1 Performance as a function of SNR

We ran all six methods—MDS1, MDS2, NMDS, PM1, PM2 and ChromSDE—on the 170 simulated datasets with varying SNR levels. Our goal here is to assess how well the different methods manage to reconstruct a known 3D structure from simulated data at different SNR levels. Remember that SNR estimates how far the empirical counts differ from their expectations; in real Hi-C data, SNR typically decreases when we have fewer reads in total, or when we want to increase the resolution of the structure. In this first series of experiments, we provide the correct count-to-distance or distance-to-count transfer functions to the methods that need them (MDS1, MDS2, PM1). In this setting, for infinite SNR, all methods should consistently estimate the correct structure.

Figure 1 shows the performance of the different methods in terms of RMSD (top) and distance error (middle) as a function of the β parameter, which controls the SNR (bottom). As expected, all methods perform well when the SNR is high, but exhibit marked differences in performance for finite SNR. In the low SNR setting (SNR < 2), both PM1 and PM2 significantly outperform all MDS-based methods, in both RMSD and distance error. Interestingly, we observe no significant difference between PM1 and PM2, which shows that there is no price to pay in terms of inferred structure if we don't specify the exponent of the distance-to-count transfer function. In this setting, PM2 is able to estimate the structure accurately enough to produce a structure of the same quality as PM1. Among MDS-based methods, we see that NMDS generally outperforms MDS2, which itself outperforms MDS1. This observation highlights that in the non-asymptotic, low SNR setting, the choice of stress function influences the performance of MDS. ChromSDE performs better than other MDS-based methods on datasets with a low SNR, corresponding to datasets with low coverage and, consequently, many non-interacting pairs of beads. This may be due to the way ChromSDE explicitly handles such pairs. On the other hand, in a more favorable setting (SNR > 2), ChromSDE does not perform as well as other MDS-based method; we hypothesize that when the coverage is high enough, taking into account non-interacting pairs of beads does not add any additional information. Since ChromSDE is not better than other MDS-based methods, and requires much longer to run, we do not report its performance on the next experiments and instead focus on the differences between the other MDS-based methods and the PM methods.

3.1.2 Metric versus nonmetric methods: robustness to incorrect parameter estimation

Three of the methods tested, which we collectively refer to as *metric* methods, require as input a count-to-distance or distance-to-count transfer function: MDS1, MDS2 and PM1. In reality, however, the DNA may not follow the ideal physical laws underlying the default transfer function

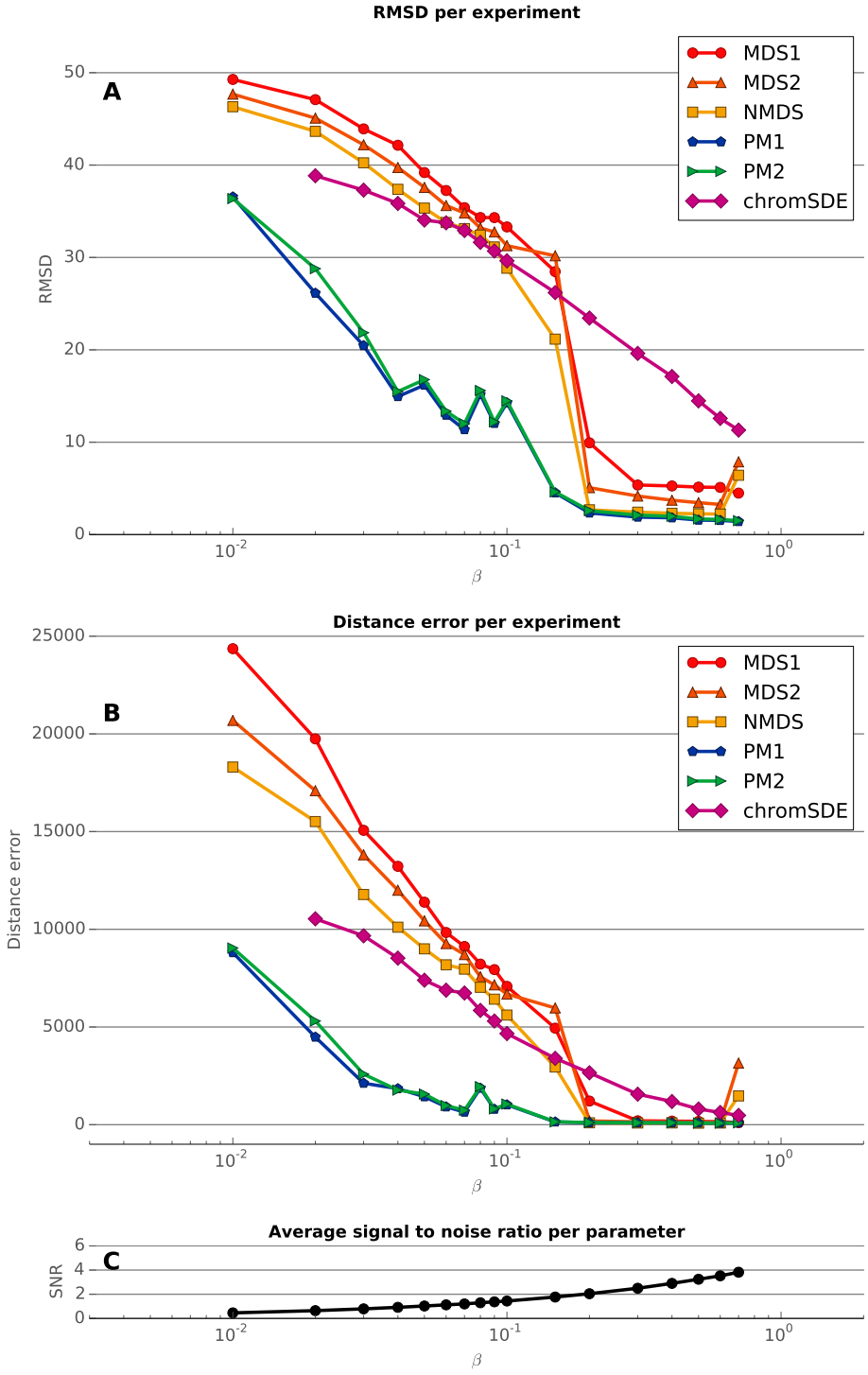

Figure 1: **Performance evaluation on simulated data, varying the parameter β. A** RMSD of each experiment for varying values of the parameter β. ChromSDE failed to yield consistent results for 14 experiments (It reported the wrong number of beads in the results file.), and the PM2 algorithm failed to converge at the desired precision for one experiment (It exceeded the maximum number of iterations.). **B** Distance error of each experiment for varying values of β. **C** Average SNR for each β. Higher SNR corresponds to better quality data.

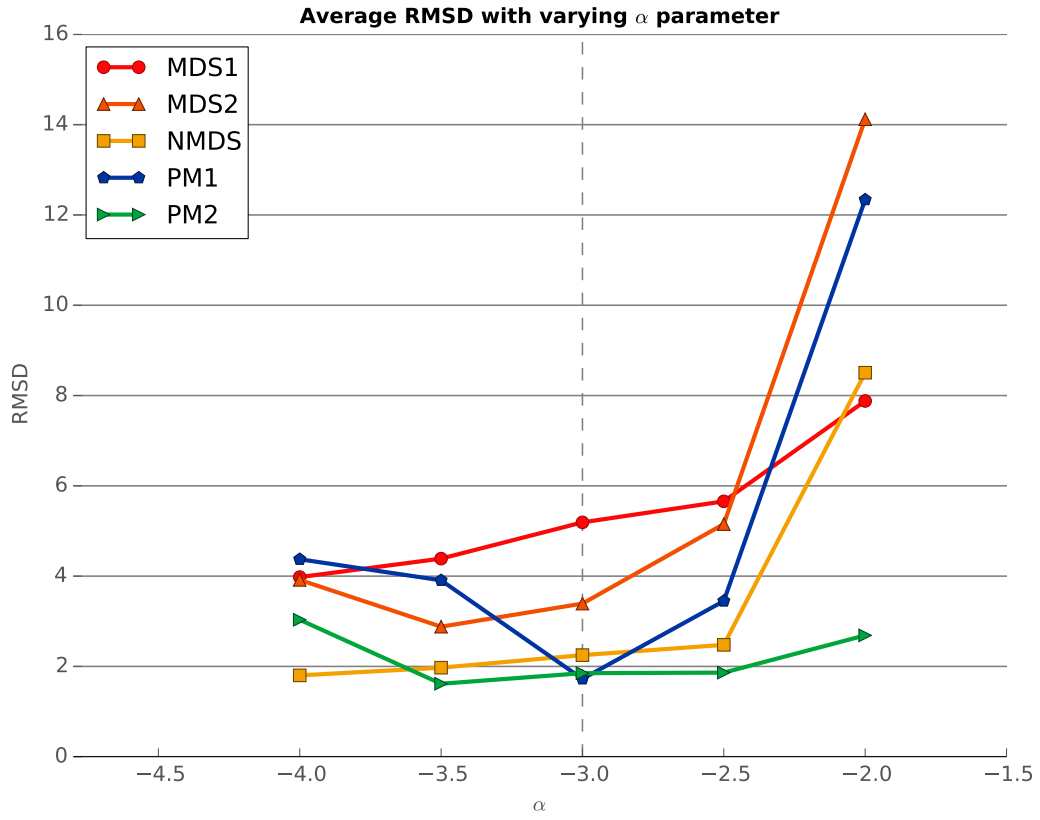

Figure 2: **Performance evaluation for simulated data, varying the parameter α.** The figure plots the average RMSD of the inferred structures for a range of α values. As α increases, the SNR of the dataset also increases.

Resolution	Beads	Corr	MDS1		MDS2		NMDS		PM1		PM2	
			RMSD	Corr	RMSD	Corr	RMSD	Corr	RMSD	Corr	RMSD	Corr
1 Mb	198	0.981	13.13	0.945	5.54	0.964	5.80	0.965	7.28	0.931	**4.92**	**0.976**
500 kb	385	0.959	10.00	0.942	5.68	0.959	5.67	0.959	7.14	0.913	**4.66**	**0.968**
200 kb	986	0.845	5.64	0.940	3.74	0.945	3.73	0.946	4.01	0.891	**3.42**	**0.958**
100 kb	1972	0.605	5.07	0.736	2.53	0.676	2.52	0.666	**2.51**	0.664	2.76	**0.771**

Table 1: **Stability across enzyme replicates.** For each resolution, the table lists the total number of beads in the inferred structure, the Spearman correlation between the two enzyme replicate datasets, and, for each inference method, the average RMSD and Spearman correlation between pairs of structures inferred from the two datasets. Boldface values correspond to the best RMSD or correlation values among all five methods. In general, higher resolution leads to a lower correlation between pairs of inferred structures.

	MDS1	MDS2	NMDS	PM1	PM2
RMSD	14.86	12.92	12.98	13.03	**11.48**
Correlation	0.781	0.754	0.738	0.737	**0.807**

Table 2: **Stability across resolution.** The table lists the average RMSD and Spearman correlation between pairs of structures of different resolutions. In bold are the lowest average RMSD and highest average Spearman correlation. These values were computed on mouse ESC HindIII libraries Dixon et al. (2012))

discussed in Section 2.4, and the structures inferred from these methods may diverge from the correct one because of miss-specification of the transfer function.

To assess this phenomenon, and evaluate the robustness of the different methods (including NMDS and PM2, which automatically infer a transfer function), we now study the performance of the methods on datasets generated with varying α parameters. We therefore run the MDS1, MDS2, NMDS, PM1 and PM2 methods on the second ensemble of simulated datasets. We provide the default transfer function to all metric methods, thus inducing a miss-specification for all simulated datasets with $\alpha \neq -3$.

Figure 2 shows the RMSD of each method, averaged over the datasets with different β, as a function of α. The performance curve of PM1, which is the best method when the data are simulated with the correct parameter $\alpha = -3$, exhibits a characteristic U-shape centered around $\alpha = -3$. This curve confirms that PM1 performs better when given the true parameter and performs worse as α moves away from -3. On the other hand, the performance curves of the two other metric methods, MDS1 and MDS2, do not exactly follow this trend: MDS1 and NMDS perform increasingly better when α decreases, and MDS2 achieves the best performance when $\alpha = -3.5$. This phenomenon occurs because in our simulation, when α decreases, the SNR for a given β increases, counterbalancing the negative effect of the transfer function miss-specification. Thus, for MDS-based methods, it is apparently more important to have more data than to have a correct α parameter. Finally, we see that, as expected, the non-metric approaches, NMDS and PM2, are more robust to transfer function misspecification than the metric approaches, because they automatically estimate it. When the parameter is wrong, PM2 outperforms the other methods for low SNR, whereas for high SNR, NMDS performs better.

3.2 Real Hi-C data

We now test the different methods on real Hi-C data. Since in this case the true consensus structure is unknown, we investigate the behaviors of the different methods in terms of their ability to infer consistent structures from different datasets and across resolutions.

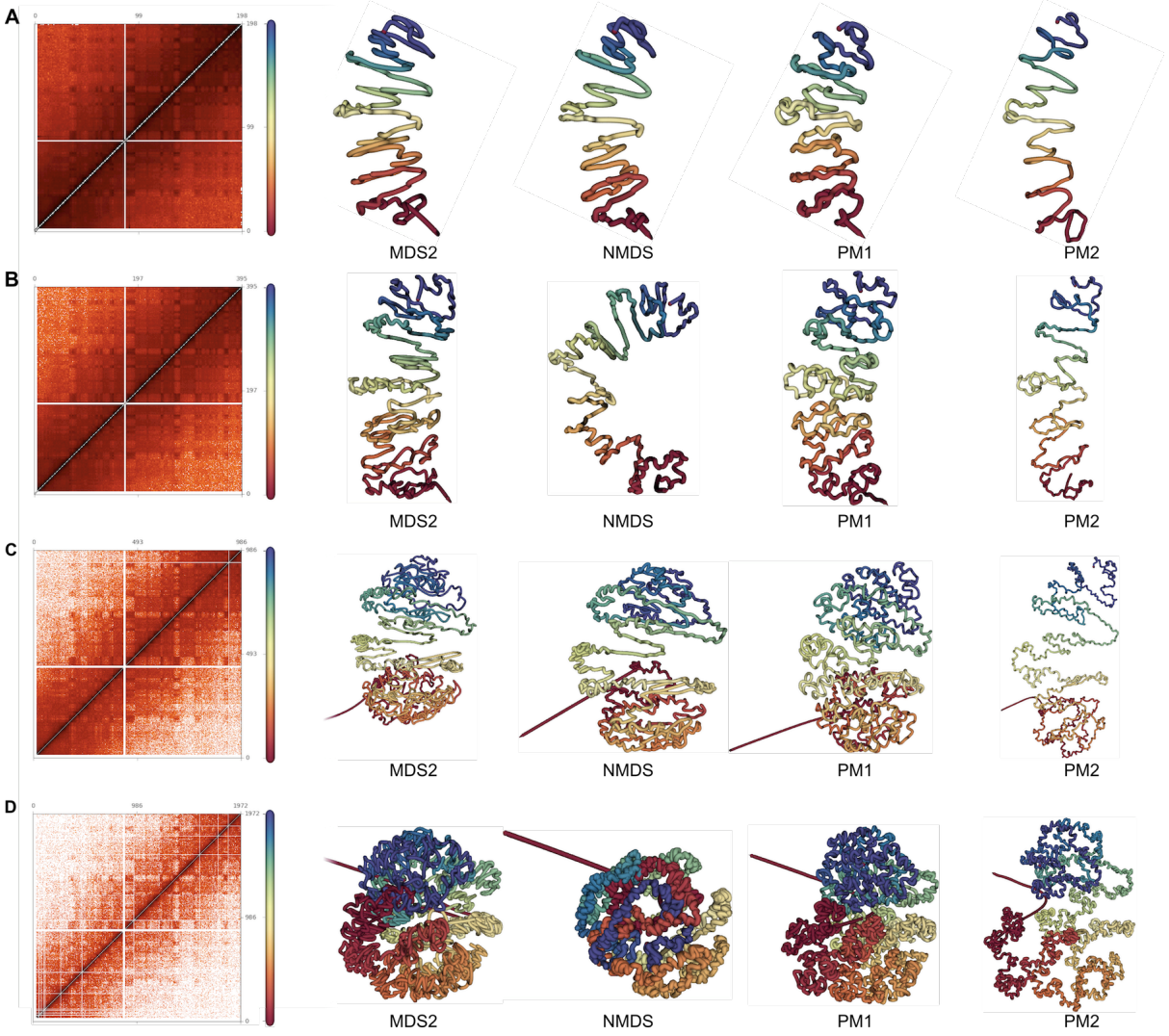

Figure 3: **Predicted structures at different resolution** Contact counts matrices and predicted structures for the MDS2, NMDS, PM1 and PM2 methods at 1 Mb (**A**), 500 kb (**B**), 200 kb (**C**), 100 kb (**D**)

3.2.1 Stability to enzyme replicates

The Hi-C assay depends upon a restriction enzyme to cleave the DNA after cross-linking, and the same sequence library can be analyzed multiple times using different enzymes. Although the resulting restriction fragments will differ, we expect *a priori* that the overall genome architecture should be the same from such replicate experiments. We therefore evaluate each genome architecture inference method with respect to the similarity of the structures inferred from two replicate Hi-C experiments that differ only in the choice of restriction enzyme. Specifically, we apply each method to two enzyme replicates, HindIII and NcoI, carried out in mouse ES cells (Dixon et al., 2012), for chromosomes 1–19.

To measure the stability of the methods, we compute (1) the Spearman correlation between the two pairwise Euclidean distance matrices of the pairs of predicted structures and (2) the RMSD between the rescaled predicted structures. Note that, before computing our two error measures, we filter out from the pair of structures any beads for which the inference hasn't been done on either dataset, i.e., beads that have zero contact counts in either data set.

To give a sense of how similar the two replicate datasets are, we also compute the Spearman correlation directly on the data, rather than on the inferred structures. As expected (Table 1), the higher the resolution is, the lower the correlation between the pairs of datasets is and the more different the inferred structures are. Across different enzyme replicates, the PM2 method yielded significantly higher correlation than all of the other methods ($p < 0.05$, signed-rank test adjusted for multiple tests with a Bonferroni correction).

3.2.2 Stability to resolution

Zhang et al. (2012) show that the mapping from contact counts to physical distance differs from one resolution to another, underscoring the importance of good parameter estimation. To study the stability of the structure inference methods to changes in resolution, we compute the RMSD between pairs of structures inferred at different resolutions. Let $(\mathbf{X}, \mathbf{Y}) \in (R^{3\times n}, R^{3\times m})$ be a pair of predicted structures such that $n < m$ (i.e., $\mathbf{X}$ is a structure at a lower resolution than $\mathbf{Y}$). We compute a downsampled structure $\mathbf{Y}^* \in X^{3\times n}$ at the same resolution as $\mathbf{X}$ by averaging the coordinates of beads. We then compute the RMSD between this new structure $\mathbf{Y}^*$ and $\mathbf{X}$, as well as a corresponding Spearman correlation of the distance matrices.

Results are shown in Figure 3 and Table 2. PM2 is significantly ($p < 0.05$) more stable to resolution changes, both in terms of RMSD and of correlation of distances.

4 Discussion and conclusion

In this work, we present a novel method for inferring a consensus genomic 3D structure from Hi-C data. The method maximizes a likelihood derived from a statistical model of the relationship between the contact counts and physical distances, and includes an automatic tuning of the parameters defining the link between a 3D distance and the Poisson parameter of the corresponding contact count. We showed in simulations that the new method outperforms a panel of MDS-based approaches, including ChromSDE, which optimize an often ad-hoc stress function. The improvement is particularly important at low SNR, corresponding to more difficult problems where we want to increase the resolution of the model with a fixed total number of reads; this is typically the situation where one expects a correct maximum likelihood estimator to outperform more *ad hoc* estimators. We also showed that misspecification in the count-to-distance transfer function can harm the performance of metric methods, while our model can adapt to unknown distributions within a parametric family. Finally, we also demonstrated, on real Hi-C data, the robustness of our methods to resolution change and enzyme duplicated datasets.

Our probabilistic model of reads is similar to the model proposed by Hu et al. (2013); however, instead of generating a family of structures by MCMC we use the model for direct

maximum likelihood estimation of a consensus structure. Although the consensus structure might not be a definitive structure *in vivo*, it provides us with a rich model for further analysis, conserving hallmarks of genome organization such as the water lily form of the budding yeast (Duan et al., 2010) or topological domains (Kalhor et al., 2011).

The Poisson model underlying our approach remains very basic and could be subject to many improvements. For example, physical constraints, such as the size of the nucleus, could be incorporated into the model. Better models for zero entries may be possible, because those can either come either from non-interacting loci or from measurement errors due to, e.g., mappability problems. Overall, expressing the structure inference problem as a maximum likelihood problem offers a principled way to improve the method by improving the probabilistic model of measured data. Inferring diploid structures, in particular, may benefit from such a modeling strategy and constitutes an exciting direction for future research.

5 Acknowledgement

We thank Nicolas Servant for insightful discussions. This work was supported by the European Research Council [SMAC-ERC-280032 to J-P.V., N.V.]; the European Commission [HEALTH-F5-2012-305626 to J-P.V., N.V.]; the French National Research Agency [ANR-11-BINF-0001 to J-P.V., N.V.] and by National Institutes of Health awards R01 AI106775, P41 GM103533 and U41 HG007000.

Bibliography

D. Bau, A. Sanyal, B. R. Lajoie, E. Capriotti, M. Byron, J. B. Lawrence, J. Dekker, and M. A. Marti-Renom. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol*, 18(1):107–114, 2011.

S. Ben-Elazar, Z. Yakhini, and I. Yanai. Spatial localization of co-regulated genes exceeds genomic gene clustering in the saccharomyces cerevisiae genome. *Nucleic Acids Res*, 41(4): 2191–2201, Feb 2013.

M. J. Best, N. Chakravarti, and V. A. Ubhaya. Minimizing separable convex functions subject to simple chain constraints. *SIAM J. on Optimization*, 10(3):658–672, July 1999. ISSN 1052-6234. doi: 10.1137/S1052623497314970. URL `http://dx.doi.org/10.1137/S1052623497314970`.

S. De and F. Michor. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat Biotechnol*, 29(12):1103–1108, 2011.

J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, 2002.

J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble. A three-dimensional model of the yeast genome. *Nature*, 465: 363–367, 2010.

G. Fudenberg and L. A. Mirny. Higher-order chromatin structure: bridging physics and biology. *Curr Opin Genet Dev.*, 22(2):115–124, 2012.

A. Y. Grosberg, S. K. Nechaev, and E. I. Shakhnovich. The role of topological constraints in the kinetics of collapse of macromolecules. *Journal de Physique*, 49(12):2095–2100, 1988.

M. Hu, K. Deng, Z. Qin, J. Dixon, S. Selvaraj, J. Fang, B. Ren, and J. S. Liu. Bayesian inference of spatial organizations of chromosomes. *PLoS Comput Biol*, 9(1):e1002893, 2013. doi: 10.1371/journal.pcbi.1002893. URL `http://dx.doi.org/10.1371/journal.pcbi.1002893`.

M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*, 9:999–1003, 2012.

R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, and L. Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol*, 30(1):90–98, 2011.

J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964. URL `http://dx.doi.org/10.1007/BF02289565`. 10.1007/BF02289565.

J. B. Kruskal and Wish. M. *Multidimensional Scaling*. Sage Publications, Beverly Hills, CA, 1977.

T. B. K. Le, M. V. Imakaev, L. A. Mirny, and M. T. Laub. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science*, 342(6159):731–734, 2013.

E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J Mach Learn Res*, 12:2825–2830, 2011.

M. Rousseau, J. Fraser, M. Ferraiuolo, J. Dostie, and M. Blanchette. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics*, 12(1):414, October 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-414. URL `http://dx.doi.org/10.1186/1471-2105-12-414`.

T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T. C. Schulz, A. J. Robins, S. Dalton, and D. M. Gilbert. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res*, 20(6):761–770, 2010.

Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488:116–120, 2012.

R. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27:125–140, 1962. ISSN 0033-3123. URL `http://dx.doi.org/10.1007/BF02289630`. 10.1007/BF02289630.

H. Tanizawa, O. Iwasaki, A. tanaka, J. R. Capizzi, P. Wickramasignhe, M. Lee, Z. Fu, and K. Noma. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res*, 38(22): 8164–8177, 2010.

H. Tjong, K. Gong, L. Chen, and F. Alber. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome Res*, 22(7):1295–1305, 2012.

B. van Steensel and J. Dekker. Genomics tools for the unraveling of chromosome architecture. *Nat Biotechnol*, 28(10):1089–1095, 2010.

A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math Program*, 106(1):25–57, May 2006. ISSN 0025-5610. doi: 10.1007/s10107-004-0559-y. URL `http://dx.doi.org/10.1007/s10107-004-0559-y`.

E. Yaffe and A. Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, 43:1059–1065, 2011.

Y. Zhang, R. P. McCord, Y. Ho, B. R. Lajoie, D. G. Hildebrand, A. C. Simon, M. S. Becker, F. W. Alt, and J. Dekker. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, 148:1–14, 2012.

Z. Zhang, G. Li, K.-C. Toh, and W.-K. Sung. Inference of spatial organizations of chromosomes using semi-definite embedding approach and Hi-C data. In M. Deng, R. Jiang, F. Sun, and X. Zhang, editors, *Proceedings of the 17th International Conference on Research in Computational Molecular Biology*, volume 7821 of *Lecture Notes in Computer Science*, pages 317–332, Berlin, Heidelberg, 2013. Springer-Verlag.